MIXTURE-OF-LORA: MODALITY-SPECIALIZED LOW-RANK ADAPTERS FOR MULTIMODAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

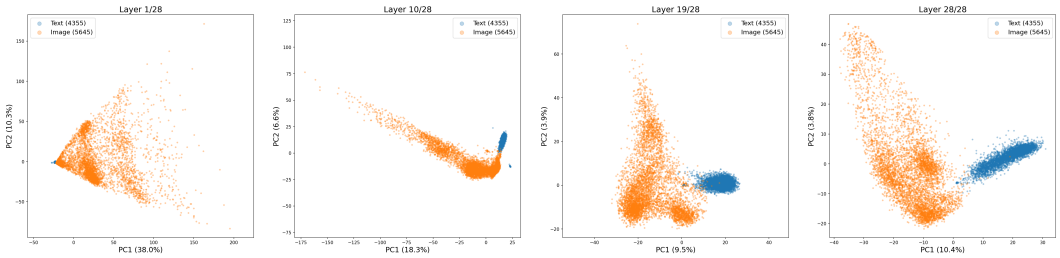
Adapting pretrained Large Language Models (LLMs) to process multimodal inputs typically requires either costly full fine-tuning or training from scratch on trillions of tokens. We observe that even when a shared low-rank adapter is used for multimodal fine-tuning, latent representations cluster by modality across layers—suggesting that modality-specific adaptation is a natural inductive bias. Motivated by this finding, we introduce Mixture-of-LoRA (MoL), a parameter-efficient fine-tuning framework that injects per-modality LoRA adapters into the frozen attention layers of a pretrained LLM, while keeping global self-attention intact for cross-modal fusion. MoL brings the modality-specialized design of Mixture-of-Transformers (MoT) to the adapter setting, enabling multimodal understanding without modifying the base model's weights. We evaluate MoL on image-text (21 benchmarks) and audio-text tasks, comparing against shared LoRA, DoRA, modality-only adapters, and their mixture variants. MoL and its DoRA counterpart (MoD) consistently outperform shared adapters—achieving the highest average scores with especially large gains on counting, spatial reasoning, and chart understanding tasks. Our analysis further reveals that per-modality adapters produce sharper attention patterns and stronger cross-modal attention compared to shared adapters.

1 INTRODUCTION

Large language models (LLMs) (Touvron et al., 2023; Abdin et al., 2024; Yang et al., 2024; OpenAI et al., 2024; Grattafiori et al., 2024; DeepSeek-AI et al., 2025) have received increasing attention from both researchers and practitioners due to their capabilities that have expanded beyond text modality. For instance, since the rise of decoder-only models originally proposed for natural language (Radford et al., 2019), many models can now process diverse modalities, e.g, text, audio, image, or videos (Bai et al., 2023; 2025; Liu et al., 2025; Kimi Team, 2026; DeepSeek-AI, 2026).

Two approaches to make LLMs multimodal have generally been considered. A first approach is to use a pretrained text-only LLM as a base model and *fine-tune* it to accept multimodal inputs. This usually involves adding an adapter module that maps the tokens from other modalities into the LLM's representation space. These approaches are particularly interesting because they leverage the vast amount of learned information already present from their extensive text-based training. Also, those approaches are cost-efficient as they avoid retraining a model from scratch, which has been known to induce significant computational costs (Liang et al., 2025). However, this line of approach usually requires carefully curating both the modalities' feature extractors and mapping modules. This often involves producing different mapping modules for each modality, thus making hyperparameter optimization more complex. Some recent work (Laurençon et al., 2024) investigated the impact of model architectures, in particular connector modules, on performance in the context of Vision-Language Models. Similarly, Verdini et al. (2025) demonstrates that the ideal adapter and feature extractor depend on the target task for speech-text models. Additionally, previous work (Das et al., 2024; Thimonier et al., 2025) highlights the sensitivity of these approaches to the training curricula. These findings emphasize the complexity of fusing new modalities into pretrained LLMs. The second type of approaches includes training foundational models from scratch to take multimodal inputs. Some of these approaches also require using pretrained vision and audio encoders (Tong et al., 2026;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Layer 1/28 (b) Layer 10/28 (c) Layer 19/28 (d) Layer 28/28

Figure 1: **Feature space analysis**. Principal Component Analysis of latent feature space for Llama3.2-3B with a shared LoRA adapter for image and text tokens across layers 1, 10, 19, and 28. Similar to dense transformers, LoRA-augmented models exhibit distinct clustering for each modality, without any modality-specific prior being instilled in the model. Results shown here are from 32 samples taken at random from the training set.

Grattafiori et al., 2024; Yang et al., 2025). However, recent work has demonstrated that multimodal models can also directly operate in the pixel space and do not necessarily need a pretrained modality encoder (Liu et al., 2026; Diao et al., 2024). While these approaches perform better than fine-tuned models across a wide range of multimodal tasks, they incur a high computational cost during training. For instance, models like Chameleon (Team, 2024) require 1.5 trillion text-image tokens, 2.9 trillion text-only tokens, and 400 billion interleaved tokens.

Recently, Liang et al. (2025) have proposed a novel architecture for training multimodal foundational models: Mixture-of-Transformers (MoT). Their proposed approach is motivated by the finding that multimodal foundational models exhibit modality-specific clustering across layers. This is also observed when fine-tuning using shared adapters as shown in Fig. 1 and 4. In short, their approach disentangles the different modalities in the token sequence within each attention layer. All modalities are processed independently to produce the query, key, value, and output matrices. Notably, the obtained modality-specific representations are concatenated (in their original order), and self-attention is applied to the concatenated representations. We extensively discuss this approach in section 3.2. While significantly more efficient than existing methods, as the required flops for the same performance decrease, this approach still requires retraining a model from scratch with trillions of multimodal tokens.

We build on this approach and propose a novel multimodal fine-tuning approach of LLMs, **Mixture-of-LoRA (MoL)**, that leverages per-modality LoRA adapters (Hu et al., 2022). In our proposed approach, we freeze the weights of a pretrained text-only LLM, in particular the query, key, value, and output matrices, and add per-modality LoRA adapters on top (see Fig. 2). Instead of learning an entire weight matrix for each modality at each attention head and multi-attention layer, we rely on efficient fine-tuning using low-rank matrices, significantly reducing overall training cost.

We evaluate our fine-tuning approach on a 3B LLM and restrict our experiments to the supervised fine-tuning stage of this multimodal setting. We start from a text-only pretrained base-LLM and experiment on two settings:

- Image-text instruct fine-tuning.
- Audio-text instruct fine-tuning.

We compare MoL adapters to different approaches to fine-tuning the model for the same compute budget. We compare to (i) standard LoRA adapter fine-tuning and (ii) image/audio-only LoRA adapters. For completeness, we also evaluate DoRA adapters and MoD (Mixture-of-DoRA). Overall, our experiments demonstrate the superiority of MoL and MoD adapters to other tested methods for the same trainable parameter count on standard multimodal benchmarks.

2 Related works

Multimodal LLMs While the first LLMs focused solely on natural languages, a wide range of multimodal foundation models has been proposed in the literature. Multimodality in foundation models initially focused on multimodal *understanding* and relied on modality-specific feature-extraction and mapping modules. Traditionally, images are encoded in the LLMs' representation space using carefully curated mapping modules, using late-fusion techniques (Alayrac et al., 2022; Chen et al., 2022; Liu et al., 2023c). Some recent work (Vallaeys et al., 2025) has investigated the optimal adapter for mapping the audio/image feature representations to the LLM representation space. As discussed in section 1, these approaches are often less costly to train than fully-trained multimodal models as they rely on pre-trained LLM backbones and require fewer tokens to fine-tune.

The other line of approaches that require training from scratch a foundational model, e.g., MoT (Liang et al., 2025), Chameleon (Team, 2024), Unified-IO (Lu et al., 2022), CM3 (Aghajanyan et al., 2022), CM3Leon (Yu et al., 2023), Emu3 Wang et al. (2024a) or Tuna-2 (Liu et al., 2026), enable both visual *understanding* and *generation*. Some of these approach require visual and audio modalities to be tokenized using a discrete dictionary from pre-trained models that involve fixed codebooks (van den Oord et al., 2017; Razavi et al., 2019; Esser et al., 2020; Liu et al., 2023a). These approaches thus allow auto-regressive generation of other modalities than just text, as the LLM's vocabulary can be extended to include these new modality-specific tokens. Other approaches rely on the transfusion (Zhou et al., 2025) framework and can rely on continuous representations of the multimodal tokens.

LLM fine-tuning A prevalent method for adapting foundation models is fine-tuning. This process specializes a general-purpose model by continuing its training on a comparatively small, task-specific dataset, enabling its application to specific domains of interest (Devlin et al., 2019). Initial approaches to parameter-efficient fine-tuning (PEFT) inserted learnable adapter modules between a model's frozen layers (Rebuffi et al., 2017; Houlsby et al., 2019; Lin et al., 2020). LoRA (Hu et al., 2022) improved this by decomposing the weight update matrix of a layer into two trainable low-rank matrices, which are learned in parallel to the frozen original weights. Subsequent work, such as DoRA (Liu et al., 2024) and LoRA+ (Hayou et al., 2024), improves the vanilla LoRA framework by respectively refining how low-rank matrices are constructed and tuning the learning-rate schedule. Close in spirit to our proposed method, MoLE (Wu et al., 2024) also leverages several adapters per layer combined with a learned gating function, enabling different inputs to activate different adapter subsets. Later work, D-MoLE (Ge et al., 2025), introduces a dynamic layer-wise expert allocator that facilitates knowledge sharing among experts. Related work includes LoRAMoE (Dou et al., 2024), which decouples world knowledge preservation from downstream task adaptation using a router network. Finally, MoLA (Gao et al., 2025), which uses a varying number of LoRA experts per layer, empirically shows that higher layers benefit from more LoRA experts. While some of these approaches have proposed using different adapters per-layer in unimodal and multimodal frameworks, our proposed approach is the first to leverage per-modality adapters for multimodal fine-tuning of pretrained LLMs while being agnostic to the adapter type, as shown in our experiments.

3 Method

In the present section, we briefly overview the vanilla attention (Vaswani et al., 2017) mechanism, then we present the mechanisms underlying the MoT architecture (Liang et al., 2025), and then discuss MoL in light of this.

3.1 Vanilla Attention

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the input token sequence, where $\mathbf{x}_i$ belongs to a modality m_i where $m_i \in \mathcal{M} = \{text, image, audio\}$. A typical transformer layer consists of the following,

$$\begin{aligned} a &= \text{Attn}\left(\mathbf{X}, \theta_{\text{attn}}\right) \\ \mathbf{h} &= \mathbf{X} + \text{LayerNorm}_{\text{attn}}(a) \\ \text{output} &= \mathbf{h} + \text{LayerNorm}_{\text{ffn}}\left(\text{FFN}(\mathbf{h}, \theta_{\text{ffn}})\right) \end{aligned} \tag{1}$$

where $\text{Attn}()$ refers to the usual self-attention mechanism. The tokens of each modality in $\mathbf{X}$ are processed altogether at each attention layer.

3.2 MoT Attention

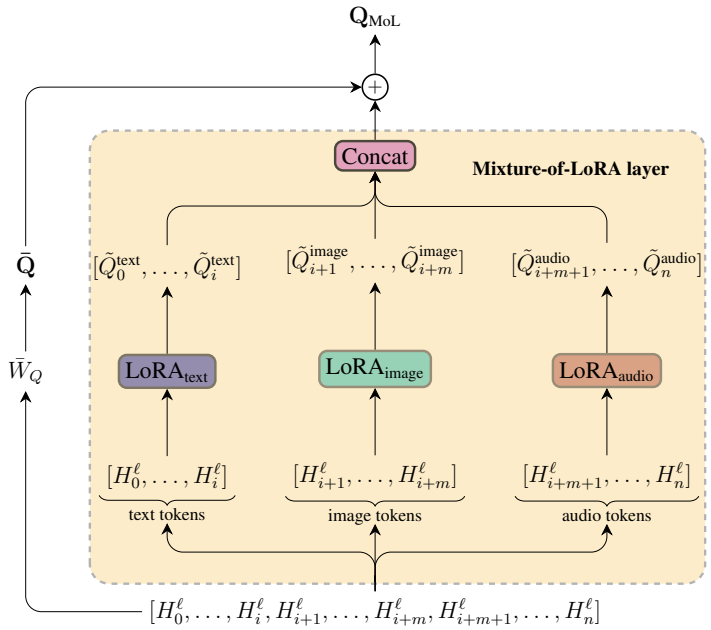


Figure 2: **Mixture-of-LoRA Layer**. Mixture-of-LoRA (MoL) layer for the query generation of an attention head at layer ℓ. For simplicity, we omit the head indices. The weights of the pretrained LLM for the corresponding layer, $\bar{W}_Q$, are frozen and produce the query matrix $\bar{\mathbf{Q}}$. The MoL layer first disentangles the tokens corresponding to each modality $m \in \mathcal{M}$ and processes them independently using modality-specific LoRA adapters. The obtained query matrices, $\{\tilde{\mathbf{Q}}^m\}_{m\in\mathcal{M}}$, are then concatenated in their original order to match the original query matrix $\mathbf{Q}$ dimension. As in the standard LoRA setting (Hu et al., 2022), $\tilde{\mathbf{Q}}$ is scaled by $\frac{\alpha}{r}$ and added to $\bar{\mathbf{Q}}$ to obtain the final query representation of the token sequence $\mathbf{Q}_{\text{MoL}}$.

Contrary to the vanilla attention setting, where each token in the sequence is processed using the same weight matrix, in the MoT approach, as proposed in Liang et al. (2025), parameters are decoupled across modalities. However, the self-attention operation is still performed on the whole token sequence. Formally, for each $m \in \mathcal{M}$, the attention layer is equipped with dedicated projection matrices, W_Q^m, W_K^m, W_V^m, that are used to process each modality *independently* of the others to obtain the query, keys, and value matrices. Let $\mathbf{x}$ be decomposed as follows,

$$\mathbf{X} = \{\underbrace{\mathbf{x}_1, \ldots, \mathbf{x}_t}_{\text{text}}, \underbrace{\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+k}}_{\text{image}}, \underbrace{\mathbf{x}_{t+k+1}, \ldots, \mathbf{x}_n}_{\text{audio}}\} \tag{2}$$

where each modality is located after the other. Note that the following also generalizes to interleaved situations. Then $\mathbf{x}_{1:t}$ is processed using $W_Q^{text}, W_K^{text}, W_V^{text}$, $\mathbf{x}_{t+1:t+k}$ is processed using $W_Q^{image}, W_K^{image}, W_V^{image}$ and $\mathbf{x}_{t+k+1:n}$ is processed using $W_Q^{audio}, W_K^{audio}, W_V^{audio}$. One then obtains,

$$\begin{aligned} Q &= \texttt{Concat}([Q^{text}, Q^{image}, Q^{audio}]) \\ K &= \texttt{Concat}([K^{text}, K^{image}, K^{audio}]) \\ V &= \texttt{Concat}([V^{text}, V^{image}, V^{audio}]) \end{aligned} \tag{3}$$

where the original sequence order is kept. Then, self-attention is performed on the whole sequence,

$$A = \texttt{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{4}$$

Let us denote, $I_m = \{i : m_i = m\}$ and $\mathbf{X}_m = \{x_i : i \in I_m\}$, then one obtains the per-modality output as

$$O_m = A_{I_m} W_O^m. \tag{5}$$

Following Liang et al. (2025) let us denote this entire processing as,

$$\texttt{GlobalAttn}(\mathbf{x}, \{\theta^m_{attn}\}_{m\in\mathcal{M}}) = \left(\texttt{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V\right)W_O^{m_i} \tag{6}$$

Then, modality-specific LayerNorm and FFN are applied to each O_m as described in equation equation 1. For a token i, this writes,

$$\begin{aligned} a &= \texttt{GlobalAttn}\left(\mathbf{X}, \{\theta^m_{attn}\}_{m\in\mathcal{M}}\right) \\ \mathbf{h}_i &= \mathbf{x}_i + \text{LayerNorm}^{m_i}_{\text{attn}}(a_i) \\ \text{output}_i &= \mathbf{h}_i + \text{LayerNorm}^{m_i}_{\text{ffn}}\left(\text{FFN}_{m_i}(\mathbf{h}_i, \theta^{m_i}_{\text{ffn}})\right). \end{aligned} \tag{7}$$

Note that the LayerNorm$(\cdot)$ operator in equation 7 is a placeholder normalization operation that can be replaced by any other relevant learnable normalization layer, e.g. RMSNorm$(\cdot)$.

3.3 Mixture-of-LoRA (MoL)

In Fig. 2 we display an overview of the mechanisms involved in a MoL layer. Let us denote the weights of a pretrained LLM as $\bar{\theta}$. Our proposed method leverages the LLM's pretrained weight matrices to fine-tune them for multimodal inputs efficiently. Similarly to section 3.2, let $\mathbf{x}$ be decomposed as shown in equation 2.

As in the standard setting, one will obtain the usual query, key, and value representations, Q, K, V using $\bar{W}_Q, \bar{W}_K, \bar{W}_V \in \bar{\theta}$. We propose using modality specific LoRA adapters, $\{(A^m_Q, B^m_Q)\}_{m\in\mathcal{M}}$, $\{(A^m_K, B^m_K)\}_{m\in\mathcal{M}}$ and $\{(A^m_V, B^m_V)\}_{m\in\mathcal{M}}$ while keeping the original layers frozen. Let r designate the chosen rank of the LoRA adapters and d the hidden dimension of the model, then $A^m_L \in \mathbb{R}^{r\times d}, B^m_L \in \mathbb{R}^{d\times r}$ for $L \in \{K, V, Q\}$. For each modality $m \in \mathcal{M}$, we compute,

$$\begin{aligned} \tilde{Q}^m &= \mathbf{X}_m B^m_Q A^m_Q \\ \tilde{K}^m &= \mathbf{X}_m B^m_K A^m_K \\ \tilde{V}^m &= \mathbf{X}_m B^m_V A^m_V \end{aligned} \tag{8}$$

Then, one concatenates the obtained representations,

$$\begin{aligned} \tilde{Q} &= \texttt{Concat}([\tilde{Q}^{text}, \tilde{Q}^{image}, \tilde{Q}^{audio}]) \\ \tilde{K} &= \texttt{Concat}([\tilde{K}^{text}, \tilde{K}^{image}, \tilde{K}^{audio}]) \\ \tilde{V} &= \texttt{Concat}([\tilde{V}^{text}, \tilde{V}^{image}, \tilde{V}^{audio}]) \end{aligned} \tag{9}$$

Those representations are then scaled and added to the representations obtained from the frozen weights of the LLM similarly to the standard LoRA setting. Formally, one computes $\mathbf{Q}_{\text{MoL}}, \mathbf{K}_{\text{MoL}}, \mathbf{V}_{\text{MoL}}$ as follows,

$$\begin{aligned} \mathbf{Q}_{\text{MoL}} &= \bar{Q} + \frac{\alpha}{r}\tilde{Q} \\ \mathbf{K}_{\text{MoL}} &= \bar{K} + \frac{\alpha}{r}\tilde{K} \\ \mathbf{V}_{\text{MoL}} &= \bar{V} + \frac{\alpha}{r}\tilde{V} \end{aligned} \tag{10}$$

Note that a similar process is performed on the output matrix, W_O. One can also include per-modality LayerNorm and LoRA adapters to the FFN networks. While including per-modality LayerNorm does not induce any significant computational overhead, replacing the pretrained FFN network with a per-modality module, as done in (Liang et al., 2025), would require retraining a significant share of the parameters of the LLM. See Algorithm 1 (Appendix) for a description of the overall process.

3.4 Input representation and tokenization

We unify all modalities into a common sequence representation to enable the LLM to process visual and audio information. We first encode images and audio into token sequences using pretrained encoders. These tokens are subsequently mapped to the LLM's representation space using dedicated

learnable modules. For image, we use an in-house vision encoder trained in a framework close to (Tschannen et al., 2023) while we rely on Whisper Large (Radford et al., 2023) to extract audio tokens. Additionally, we include special tokens delimiting a modality's tokens in the LLM's vocabulary. We rely on this approach as it is the standard approach to enabling LLMs to also input other modalities than text. We also include `[IMG BREAK]` tokens to enable 2D position understanding of the model.

4 EXPERIMENTS

In this section, we first discuss the experimental settings, including hyperparameter settings and training datasets, and then present the results.

4.1 EXPERIMENTAL SETTINGS

Datasets For both settings, image-text and audio-text, we rely on a curated selection of in-house multimodal SFT datasets. We restrict ourselves to instruct multimodal datasets for both setups. For the image-text setup, we sample so that image tokens account for approximately 80% of the total tokens seen by the LLM. The rationale is that, because we rely on a base model that has already seen a considerable number of text tokens during the pretraining phase, the fine-tuning phase should primarily focus on the new modality.

Adapters We compare five adapter setups for fine-tuning the pretrained text-only LLM: a shared LoRA (Hu et al., 2022) and DoRA (Liu et al., 2024) adapter, their mixture equivalents, Mixture-of-LoRA (MoL) and Mixture-of-DoRA (MoD), where each modality is assigned its own adapter, and an image-only LoRA baseline, where only image tokens receive adapters while text tokens pass through the frozen LLM unchanged. For both the image-text and audio-text setups, the modality-agnostic baselines (LoRA, DoRA) use a single shared adapter with rank 32 and $\alpha = 64$. MoL and MoD employ per-modality adapters with rank 16 and $\alpha = 16$ for text, and rank 16 and $\alpha = 32$ for image/audio, along with per-modality RMSNorm layers. We select the rank of the LoRA and DoRA baselines so as to match the trainable parameter count of MoL and MoD.

Model hyperparameters Hyperparameters detailed in the paragraph are shared across all adapter setups. We conduct our experiments using Llama3.2-3B (Grattafiori et al., 2024) as the base LLM, which was pretrained on text-only tokens. In both audio–text and image–text settings, the LLM and the modality encoder are frozen; only the MoL adapters, the multimodal projector, and per-modality RMSNorm layers are trained. We fine-tune the model by augmenting its attention layers with MoL, MoD, LoRA, and DoRA adapters on the Q, K, V, and O projections. For MoD and MoL, we also enable per-modality RMSNorm.

For the **image–text** setting, we rely on an in-house ViT-L-14 vision encoder with a maximum image size of 1540 pixels and a patch size of 14. Visual features are projected into the LLM's embedding space via a patch-merging projector with a spatial merge size of 2, preceded by an RMSNorm layer. We optimize the model's weights using AdamW with a base learning rate of $5.e^{-5}$ and weight decay of 0.05. All parameter groups follow a cosine-then-linear schedule with a minimum learning rate ratio of 0.01, except for the image MoL adapters, which use a minimum ratio of 0.05. Warm-up periods are set to 1000 steps for all parameters. We clip gradients to a max norm of 2.0 and use an effective batch size of 128 with a context length of 8192, providing the model with approximately 1M tokens per optimization step. We train for 20,000 steps on 32 H100 GPUs.

For the **audio–text** setting, we use a frozen Whisper Large v3 (Radford et al., 2023) encoder operating at 12.5 Hz with 128 mel-frequency bins. Audio features are projected into the LLM's embedding space through an MLP projector. We optimize the model's weights using AdamW with a base learning rate of $1e^{-5}$ and weight decay of 0.01. Similarly to the image-text setup, all parameter groups follow a cosine-then-linear schedule with a minimum learning rate ratio of 0.01 and warm-up period is set to 1000 steps for all groups. We clip gradients to a max norm of 1.0 as we found training to be more unstable than for the image-text setting. Effective batch size is set to 64, with a context length of 8192, providing the model with 500k tokens per optimization step. Training lasts for 20,000 steps on 32 H100 GPUs.

4.2 RESULTS

Benchmark / Metric	DoRA	LoRA	LoRA (img only)	MoL (Ours)	MoD (Ours)
Document & OCR					
DocVQA[§] (Mathew et al., 2021b)	86.42	**87.14**	85.05	85.83	84.74
InfographicVQA[†] (Mathew et al., 2021a)	**39.41**	38.63	33.70	37.31	38.84
OCRBench-v2[¶] (EN) (Fu et al., 2025)	17.22	14.03	15.96	12.16	**18.25**
TextVQA[†] (Singh et al., 2019)	74.06	73.18	56.56	70.48	**74.12**
Category Avg	**54.28**	53.25	47.82	51.45	53.99
Chart, Plot & Table					
ChartQA[*] (Masry et al., 2022)	54.72	53.32	59.96	60.72	**61.96**
ChartX[*] (Xia et al., 2025)	39.62	38.12	38.92	**42.20**	42.14
CharXiv[*] (Wang et al., 2024b)	50.99	52.53	41.76	47.25	**53.85**
DVQA[†] (Kafle et al., 2018)	70.37	72.74	54.22	**77.30**	67.39
PlotQA[*] (Methani et al., 2020)	37.37	40.05	48.63	47.94	**48.79**
TabMWP[†] (Lu et al., 2023)	24.64	23.90	**33.07**	31.63	32.84
Category Avg	46.29	46.78	46.09	**51.17**	51.16
General VQA					
GQA[†] (Hudson & Manning, 2019)	49.03	50.59	42.34	49.36	**51.24**
VQAv2[‡] (Antol et al., 2015)	73.09	73.14	62.32	72.87	**73.53**
SeedBench[†] (Ying et al., 2025)	66.15	**66.40**	50.90	64.57	65.60
Category Avg	62.76	63.38	51.85	62.27	**63.46**
Counting & Spatial					
CountBench[†] (Paiss et al., 2023)	25.99	33.73	19.44	**61.31**	50.40
TallyQA[†] (Acharya et al., 2018)	76.59	**76.77**	75.48	72.07	75.42
VSR[†] (Liu et al., 2023b)	58.82	**60.00**	53.58	55.76	54.35
CVBench[†] (Zhu et al., 2026)	**68.01**	67.93	39.69	57.28	58.68
Category Avg	57.35	59.61	47.05	**61.61**	59.71
Visual Reasoning					
AI2D[†] (Kembhavi et al., 2016)	86.98	86.76	84.39	86.11	**87.47**
IconQA[†] (Lu et al., 2021)	**85.15**	84.48	82.16	80.40	84.69
V-STaR[†] (Cheng et al., 2025)	**60.21**	57.07	48.17	58.12	54.97
Winoground (I2T)[‖] (Thrush et al., 2022)	18.25	15.75	21.75	**25.75**	22.25
Category Avg	**62.65**	61.02	59.12	62.60	62.35
Overall Average	55.39	55.54	49.91	56.97	**57.21**

Table 1: **Instruct Image-Text performance comparison**. We instruct fine-tune Llama3.2-3B model and compare its performance depending on the adapter type. We evaluate DoRA (Liu et al., 2024), LoRA (Hu et al., 2022), LoRA adapters on image tokens only, Mixture-of-LoRA (MoL) and Mixture-of-DoRA (MoD), with identical trainable parameter count. We observe overall that both MoL and MoD obtain the best performance in comparison to other adapters. In particular, we observe that MoD obtains the highest average metric and MoL the second highest. *Metrics*: [†]Exact match, [§]Average Normalized Levenshtein Similarity (ANLS) (Biten et al., 2019), [¶]Fuzzy match, [*]Relaxed correctness, [‡]VQA match, [‖]Winoground Accuracy.

Image-text We evaluate on 21 benchmarks spanning five categories: document understanding and OCR (DocVQA, InfographicVQA, OCRBench-v2, TextVQA), chart/plot/table analysis (ChartQA, ChartX, CharXiv, DVQA, PlotQA, TabMWP), general visual question answering (GQA, VQAv2, SeedBench), counting and spatial reasoning (CountBench, TallyQA, VSR, CVBench), and visual reasoning (AI2D, IconQA, V-STaR, Winoground). Results are reported in Table 1. We observe that both MoL and MoD achieve the best overall performance, with MoD obtaining the highest average

score (57.2) and MoL, the second-highest average (57.0). The gap is especially pronounced on counting and spatial reasoning tasks, where MoL achieves 61.6 category average compared to 59.6 for LoRA and 57.4 for DoRA—driven primarily by CountBench, where MoL (61.3) and MoD (50.4) outperform both LoRA (33.7) and DoRA (26.0). A similar pattern emerges for chart, plot, and table analysis, where MoL and MoD reach a category average of 51.2 versus 46.8 for LoRA and 46.3 for DoRA, with DVQA showing the largest gap (MoL: 77.3 vs. LoRA: 72.7). On document and OCR tasks, performance is more uniform across methods.The image-only LoRA baseline performs notably worse overall (49.9 average), confirming that adapting text token processing is important for strong multimodal performance. It remains competitive on chart and table tasks (46.1 category average), suggesting that image-specific adaptation captures much of the visual reasoning needed for structured visual understanding. These results demonstrate that per-modality adapter specialization, as provided by MoL and MoD, consistently outperforms both shared and image-only adaptation strategies.

Audio-text To evaluate audio capacities, we rely on three benchmarks ranging from audio captioning to sound classification and multiple-choice audio QA. Namely, we rely on AudioCaps (Kim et al., 2019) reformulated as an instruction-following task, VocalSound (Gong et al., 2022) reformulated as open-ended generation, and MMAU (Sakshi et al., 2025) (speech subset only).

5 Discussion

5.1 MoL adapter for non-text modalities

Given the base model's extensive text pre-training, we investigated the necessity of a dedicated text MoL adapter during multimodal fine-tuning. We display in Tab 1 the performance of a model finetuned with an image-only LoRA adapter and compare its performance to other tested methods. For a fair comparison, we select adapters ranks in order to make trainable parameters match between tested methods.

Apart from two benchmarks, PlotQA and TabMWP, the image-only LoRA setting is outperformed by all other methods including the base LoRA and DoRA settings. The overall average across all benchmarks also indicates that either sharing the adapter across modalities or having one adapter per modality yields significantly better performance than the image-only setting. We attribute this discrepancy to image adapters having to bear the burden of cross-modal alignment, requiring them to map visual features into a fixed and potentially suboptimal textual representation space. Including text adapters resolves this by introducing complementary transformation: the text adapters learn to condition the token representations for optimal fusion, thereby allowing the image adapters to focus on their primary task of modality-specific feature representation. We also display in Fig. 5 the training loss curves of a base model augmented by MoL, shared LoRA and image-only LoRA and observe how the MoL alternative obtains a significantly lower loss curve than the two alternatives.

5.2 Asymmetry of Per-Modality Adapter Magnitudes

In the image-text MoL setting, to assess whether the two modality-specific adapters learn updates of comparable scale or whether one modality demands larger corrections to the pretrained weights, we compute the ratio of the Frobenius norms of the effective low-rank updates, $\|\Delta_{\text{img}}\|_F / \|\Delta_{\text{txt}}\|_F$, where $\Delta = (\alpha/r)\, BA$, for each attention projection (Q, K, V, O) at every layer.

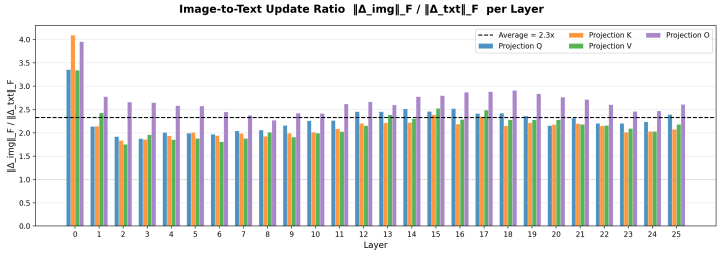


Figure 3: Ratio $\|\Delta_{\text{img}}\|_F / \|\Delta_{\text{txt}}\|_F$ per layer for each attention projection (Q, K, V, O). The dashed line marks the overall average ($2.3\times$).

Figure 3 shows that the image adapter consistently applies larger corrections than the text adapter across all layers and projections, with an overall mean ratio of $2.3\times$. This asymmetry is most pronounced in the earliest layer, where the ratio reaches 3.3–$4.1\times$ depending on the projection,

and gradually decreases toward $\sim 2\times$ in the middle layers. The pattern is consistent with the intuition that early layers must bridge a larger representational gap for visual tokens which enter the backbone through a learned projection rather than a native embedding while later layers operate on representations that have already been partially aligned by preceding blocks. Crucially, the ratio remains above 1.0 at every layer, confirming that the image modality uniformly requires a stronger adaptation signal and that this asymmetry is not concentrated in a handful of layers but is a global property of the network.

5.3 Cross-Modal Attention Patterns

We compare the attention patterns of both adapters at inference time on 16 random samples from the training set. For each query token, we compute (i) the entropy of its attention distribution, $H = -\sum_j a_j \ln a_j$ (in nats), averaged separately over text and image queries, and (ii) the cross-modal attention fraction, i.e. the share of attention mass falling on keys of the opposite modality. Table 2 reports averages across the 28 layers and all 24 heads; per-layer curves are in Figure 6 (Appendix).

Table 2: Cross-modal attention statistics, by query modality.

	Shared LoRA		MoL	
	Text	Image	Text	Image
Entropy (nats)	1.81	2.02	1.91	**1.61**
Cross-modal frac.	0.030	0.717	0.024	**0.798**

MoL produces markedly lower image query entropy (1.61 vs. 2.02 nats), indicating that its dedicated image adapter learns sharper, more selective attention patterns for visual tokens. It also routes a larger fraction of image attention toward text keys (0.80 vs. 0.72), grounding visual representations more tightly in the surrounding textual context. Text query statistics remain comparable, as expected given that text attention is largely governed by the frozen base model. Overall, per-modality routing enables more focused cross-modal integration for the minority modality without disrupting text attention.

6 Conclusion

The present work proposes Mixture-of-LoRA (MoL), a novel, parameter-efficient method for equipping pretrained large language models with multimodal capabilities. Our experiments in image-text and audio-text settings demonstrate that MoL enables LLMs to understand multimodal data effectively. MoL significantly outperforms a standard LoRA baseline with negligible computational overhead. By leveraging the extensive knowledge of the base LLM, our approach circumvents the need for training a multimodal model from scratch, thus presenting a computationally efficient alternative to existing methods (Team, 2024; Liang et al., 2025).

Limitations and future work Our current framework relies on modality-specific feature extractors that map each modality into a continuous representation space, requiring a dedicated encoder and projector per modality, and is restricted to multimodal understanding rather than generation. A promising direction for future work is to extend MoL to discrete tokenization schemes that enable autoregressive generation of non-text modalities, and to study its behavior with additional encoders such as WavLM (Chen et al., 2021) for audio or CLIP (Radford et al., 2021) for vision.

AI use statement

We used generative AI tools to (i) help with LaTeX and Ti*k*Z formatting, (ii) proofread and polish the manuscript, (iii) implement some of the ablation methods, (iv) design and give feedback on experiments, and (v) assist in interpreting results. We did not use generative AI tools to clean or reformat datasets, support qualitative or thematic data analysis, generate synthetic datasets, or propose or refine hypotheses. Developing theoretical models or conceptual frameworks, formulating mathematical claims or providing critical ingredients for proving them, and writing proofs are not applicable to this work. All AI-assisted code was thoroughly reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, and Nguyen Bach et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. *CoRR*, abs/1810.12440, 2018.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering, 2019.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18009–18019, 2022.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *CoRR*, abs/2110.13900, 2021.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning, 2025.

Nilaksh Das, Saket Dingliwal, S. Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. Speechverse: A large-scale generalizable audio language model. *ArXiv*, abs/2405.08295, 2024.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and Runxin Xu et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1932–1945, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.106.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *CoRR*, abs/2012.09841, 2020.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2025.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Ruibo Liu, Baochen Sun, Yawen Zhang, Daiyi Peng, Xiaoyuan Guo, and Vs Subrahmanian. MoLA: MoE LoRA with layer-wise expert allocation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5112–5127, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.284.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum loRA experts for continual multimodal instruction tuning. In *Forty-second International Conference on Machine Learning*, 2025.

Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155. IEEE, May 2022. doi: 10.1109/icassp43922.2022.9746828.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman et al. The llama 3 herd of models, 2024.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Forty-first International Conference on Machine Learning*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *CoRR*, abs/1902.00751, 2019.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for compositional question answering over real-world images. *CoRR*, abs/1902.09506, 2019.

Kushal Kafle, Scott Cohen, Brian L. Price, and Christopher Kanan. DVQA: understanding data visualizations via question answering. *CoRR*, abs/1801.08163, 2018.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

Kimi Team. Kimi k2.5: Visual agentic intelligence, 2026.

Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 441–459, Online, November 2020. Association for Computational Linguistics.

Alexander H. Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and James R. Glass. DinoSR: Self-distillation and online clustering for self-supervised speech representation learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devendra Singh Chaplot, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gabrielle Berrada, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jason Rute, Jean-Hadrien Chabran, Jessica Chudnovsky, Joachim Studnia, Joep Barmentlo, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Karmesh Yadav, Kartik Khandelwal, Kush Jain, Lélio Renard Lavaud, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Matthieu Dinot, Maxime Darrin, Maximilian Augustin, Mickaël Seznec, Neha Gupta, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Rémi Delacourt, Romain Sauvestre, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Shashwat Dalal, Siddharth Gandhi, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Tom Bewley, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, Yihan Wan, and Yunhao Tang. Voxtral, 2025.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c.

Shih-yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, Tao Xiang, Wenhu Chen, Ping Luo, Luke Zettlemoyer, and Yuren Cong. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation, 2026.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. *CoRR*, abs/2104.12756, 2021a.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pp. 2200–2209, 2021b.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and Aidan Clark et al. GPT-4o system card, 2024.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 506–516, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. *CoRR*, abs/1904.08920, 2019.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Hugo Thimonier, Antony Perzo, and Renaud Seguier. EmoSLLM: Parameter-efficient adaptation of LLMs for speech emotion recognition. In *39th Conference on Neural Information Processing Systems, Workshop: Multimodal Representation Learning for Healthcare*, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, Yann LeCun, and Saining Xie. Beyond language modeling: An exploration of multimodal pretraining, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Michael Tschannen, Manoj Kumar, Andreas Peter Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Théophane Vallaeys, Mustafa Shukor, Matthieu Cord, and Jakob Verbeek. Improved baselines for data-efficient perceptual augmentation of llms. In *Computer Vision – ECCV 2024 Workshops*, pp. 369–387, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-92089-9.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sébastien Bratières, Paolo Merialdo, and Simone Scardapane. How to connect speech foundation models and large language models? what matters and what does not, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024a.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024b.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, and Chengpeng Li et al. Qwen2 technical report, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,

Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jie Ying, Zihong Chen, Zhefan Wang, Wanli Jiang, Chenyang Wang, Zhonghang Yuan, Haoyang Su, Huanjun Kong, Fan Yang, and Nanqing Dong. SeedBench: A multi-task benchmark for evaluating large language models in seed science. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31395–31449, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, Xiongfei Yao, and Shuaiwei Jiao. Cvbench: Benchmarking cross-video synergies for complex multimodal reasoning, 2026.

A METHOD

Algorithm 1 Mixture-of-LoRA (MoL) Layer

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be the input sequence, where $\mathbf{x}_i \in \mathbb{R}^d$ and $m_i \in \{text, image, audio\}$ is the modality of token $\mathbf{x}_i$.
Let $\mathcal{M} = \{text, image, audio\}$ be the set of modalities.
Let $W_Q, W_K, W_V, W_O \in \bar{\theta}$ denote the frozen projection matrices of the pretrained LLM.
Let $(A_Q^m B_Q^m), (A_K^m B_K^m), (A_V^m B_V^m), (A_O^m B_O^m)$ denote the LoRA adapters for modality m, r the corresponding rank and α the scaling factor.
Let FFN denote the shared (non-modality-specific) feed-forward network and LayerNorm^m the per-modality normalization layers.
for each modality $m \in \mathcal{M}$ **do**
    $I_m \leftarrow \{i : m_i = m\}$ ▷ Indices of tokens for modality m
    $X_m \leftarrow \{\mathbf{x}_i : i \in I_m\}$ ▷ Group tokens by modality
    $\tilde{Q}^m \leftarrow \mathbf{X}_m B_Q^m A_Q^m$ ▷ Modality-specific LoRA adapters
    $\tilde{K}^m \leftarrow \mathbf{X}_m B_K^m A_K^m$
    $\tilde{V}^m \leftarrow \mathbf{X}_m B_V^m A_V^m$
end for
$\tilde{Q} \leftarrow \texttt{Concat}(\{\tilde{Q}^m\}_{m\in\mathcal{M}})$, $\tilde{K} \leftarrow \texttt{Concat}(\{\tilde{K}^m\}_{m\in\mathcal{M}})$, $\tilde{V} \leftarrow \texttt{Concat}(\{\tilde{V}^m\}_{m\in\mathcal{M}})$ ▷ Aggregate LoRA representations
$\mathbf{Q}_{\text{MoL}} \leftarrow XW_Q + \frac{\alpha}{r}\tilde{Q}$, $\mathbf{K}_{\text{MoL}} \leftarrow XW_K + \frac{\alpha}{r}\tilde{K}$, $\mathbf{V}_{\text{MoL}} \leftarrow XW_V + \frac{\alpha}{r}\tilde{V}$
$A \leftarrow \text{softmax}\left(\frac{\mathbf{Q}_{\text{MoL}}\mathbf{K}_{\text{MoL}}^\top}{\sqrt{d_k}}\right)\mathbf{V}_{\text{MoL}}$ ▷ Global self-attention
for each modality $m \in \mathcal{M}$ **do**
    $\tilde{O}^m \leftarrow A_{I_m} B_O^m A_O^m$ ▷ Modality-specific LoRA output projection
    $O^m \leftarrow A_{I_m} W_O + \frac{\alpha}{r}\tilde{O}^m$ ▷ Frozen output projection plus LoRA update
    $H^m \leftarrow X_m + \text{LayerNorm}_{\text{attn}}^m(O^m)$ ▷ Residual connection and layer norm
    $F^m \leftarrow \text{FFN}(H^m)$ ▷ Shared feed-forward network (no per-modality adapter)
    $Y^m \leftarrow H^m + \text{LayerNorm}_{\text{ffn}}^m(F^m)$ ▷ Residual connection and layer norm
end for
return $\{Y^m : m \in \mathcal{M}\}$ ▷ Return transformer layer outputs

B FEATURE SPACE ANALYSIS

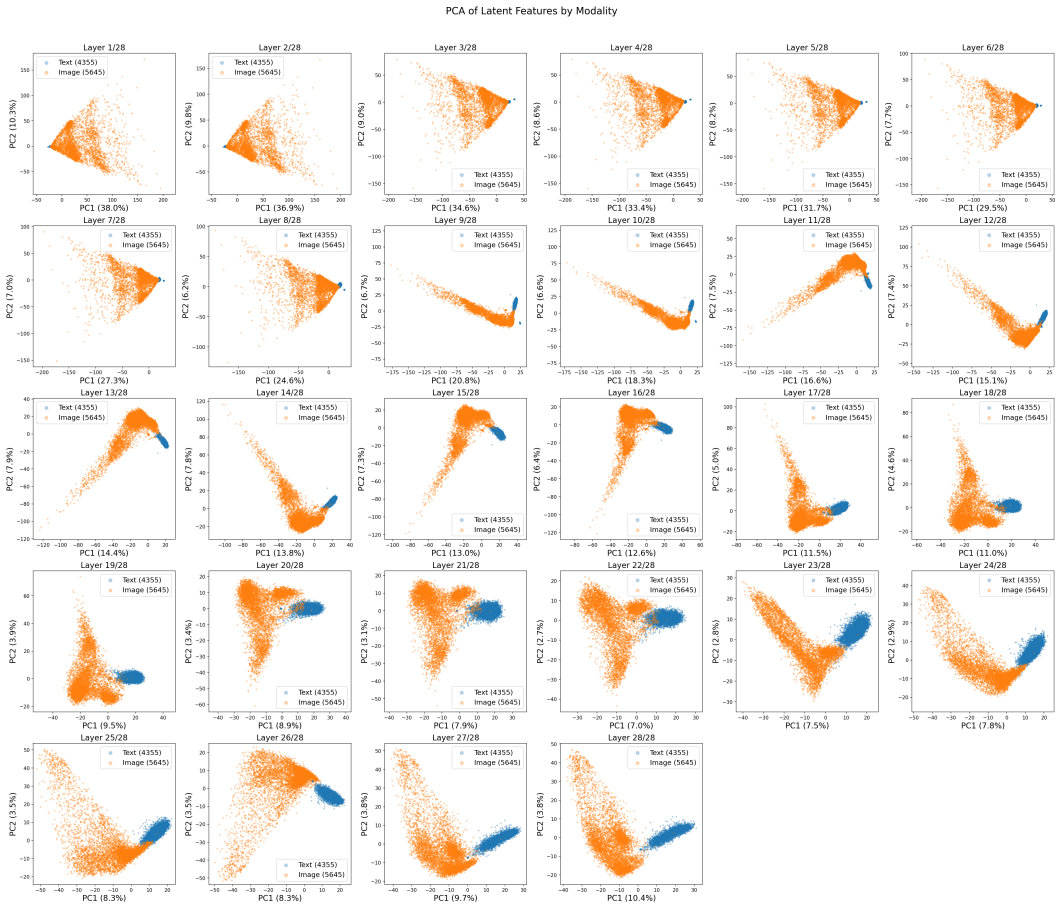


Figure 4: **Feature space analysis**. Principal Component Analysis of latent feature space for Llama3.2-3B with a shared LoRA adapter for image and text tokens across all layers. We observe distinct clustering for each modality at each layer, particularly in the latter layers of the model. The clustering particularly appears after layer 10 and is accentuated in the later layers.

C IMAGE-ONLY ABLATION

To illustrate the discrepancy in performance between shared LoRA, image-only LoRA and per-modality MoL adapters, we plot in Fig. 5 the training loss curves of all three methods.

Since the model only predicts text tokens, the training loss reported here is a text-prediction loss, with image tokens serving as conditioning context from the frozen encoder and projector. Consistent with the evaluation results in Table 1, MoL attains the lowest training loss of the three configurations, indicating that per-modality specialization eases optimization in addition to improving downstream performance. The image-only setting is the most instructive case: it reaches a lower training loss than shared LoRA yet lags noticeably at evaluation time.

Because the text *pathway* is left frozen, the image adapter must shoulder all of the cross-modal alignment on its own, and it appears to do so by shifting visual features toward a training-set-specific solution that fits the seen text distribution but generalizes poorly to the diverse question formats of the evaluation benchmarks. Shared LoRA avoids this by adapting both modalities within a single adapter, keeping the image and text pathways co-adapted and recovering most of the evaluation gap. These observations reinforce the conclusion of Section 5.1: a dedicated text adapter is necessary so that the image adapters can specialize in modality-specific feature extraction while the text adapter conditions the shared representation space for fusion.



Figure 5: **Training loss curves**. We plot the training loss curves of shared LoRA, image-only LoRA and per-modality MoL adapters in the image-text setting. We observe that while evaluation performances displayed in Tab. 1 correlates well with training loss for the MoL setting as it performs best and shows the lowest training loss, shared LoRA seem to show a higher training loss than image-only while yielding better evaluation performance.

D AUDIO-TEXT RESULTS

Table 3: **Audio evaluation results**. **AudioCaps** (audio→caption) reports corpus **BLEU** and **ROUGE** between the generated caption and human references (higher is better). **MMAU** (multiple-choice audio QA) reports the weighted-average **letter-instruction match** (exact option-letter accuracy) sliced *by difficulty* (easy/medium/hard). **VocalSound** (audio→vocalization class) reports **exact-match** (VS-EM) of the predicted label. All metrics are in $[0, 1]$, higher is better.

Benchmark	Metric	DoRA	LoRA	MoL	MoD
AudioCaps (audio→caption)	BLEU	0.065	0.059	<u>0.091</u>	**0.112**
AudioCaps (audio→caption)	ROUGE	0.189	0.186	**0.249**	0.246
MMAU (audio QA, by difficulty)	letter-instruct match	0.345	0.376	<u>0.442</u>	**0.476**
VocalSound (audio→label)	exact-match	<u>0.366</u>	0.335	0.365	**0.392**

Across all three audio benchmarks, the mixture adapters outperform their shared-adapter counterparts, mirroring the trend observed in the image-text setting. MoD achieves the highest score on AudioCaps BLEU (0.112), MMAU (0.476), and VocalSound (0.392), while MoL leads on AudioCaps ROUGE (0.249) and remains a close second on the remaining metrics. The gains are largest on MMAU, where MoD improves over LoRA by 10 absolute points in letter-instruction match, suggesting that per-modality adapters help particularly on tasks that require mapping audio inputs to structured textual answers. The smallest margin is on VocalSound, where DoRA is competitive with MoL (0.366 vs. 0.365), indicating that simple open-ended classification benefits less from modality-specific specialization than generative captioning or multiple-choice comprehension. Overall, these results confirm that the advantages of MoL and MoD extend beyond vision to the audio modality.

E COMPUTE

Table 4: Training throughput for adapter methods. Let us mention here that no attempt at optimizing throughput was done on either method, and thus reported TFLOP/s do not necessarily reflect real speed differences between models.

Method	Model TFLOPs/s	MFU	Tokens/sec (avg)	Per-step latency (s)	Overhead vs. LoRA
LoRA	527.7	0.528	14,401	2.126	—
Mixture-of-LoRA	413.1	0.413	11,245	2.722	+28.0%
DoRA	502.7	0.503	13,701	2.235	+5.1%
Mixture-of-DoRA	377.4	0.377	10,262	2.982	+40.2%

The mixture adapters incur a measurable but modest computational overhead relative to the shared-adapter baselines. MoL adds 28.0% per-step latency over LoRA, while MoD is the most expensive configuration at 40.2%, reflecting the magnitude decomposition of DoRA combined with per-modality duplication. In both cases the overhead stems from a lower model FLOP utilization (MFU of 0.413 for MoL and 0.377 for MoD versus 0.528 for LoRA) rather than from a larger theoretical compute budget, as the adapters themselves remain low-rank; the cost is dominated by the extra kernel launches and memory traffic needed to route each modality through its own projection. DoRA sits in between, with only 5.1% overhead over LoRA, showing that the magnitude decomposition itself is cheap and that the bulk of the cost in MoD comes from the mixture. Given the consistent evaluation gains reported in Tables 1 and 3, we view this overhead as a favorable trade-off for parameter-matched adapter methods.

F ADAPTER IMPACTS

We investigate which attention projection the Mixture-of-LoRA (MoL) adapters influence most for the Image-text setting. The MoL block attaches two modality adapters *image* and *text* to each of the Q, K, V, and O projections of every layer so the relative impact across projections is encoded entirely in the learned adapter magnitudes.

For each layer and projection we formed the effective scaled update $\Delta = (\alpha/r)\, BA$ for each modality-specific adapters, and summed them. We then compare their Frobenius norm to that of the frozen base weight W. Aggregated over the 28 decoder layers, the relative update magnitude $\|\Delta_{\text{img}} + \Delta_{\text{txt}}\|_F / \|W\|_F$ is largest for Q (0.84), followed by K (0.77), V (0.74), and O (0.72); the same ranking holds for the unscaled and for the mean absolute value of the update. The ordering is stable across layers: mean per-layer ratio 0.816 ± 0.152 for $\mathtt{q}$ versus 0.737 ± 0.169, 0.732 ± 0.130, 0.709 ± 0.161 for $\mathtt{k}/\mathtt{v}/\mathtt{o}$. Only the final layer behaves as an outlier. Two caveats apply: the mixture is strongly image-dominated ($\|\Delta_{\text{img}}\|_F \approx 2.3 \times \|\Delta_{\text{txt}}\|_F$ across all projections, consistent with the image adapter requiring a stronger adaptation signal than the text adapter), so the Q-dominance is driven primarily by the image adapter.

G CROSS-MODAL ATTENTION PATTERNS

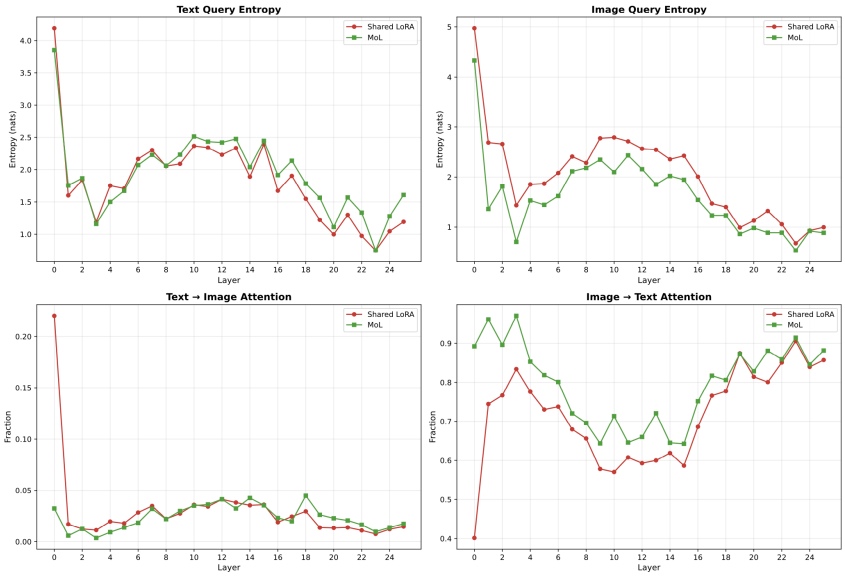


Figure 6: Per-layer cross-modal attention analysis comparing Shared LoRA and MoL adapters on Llama 3.2-3B. **Top:** attention entropy (nats) for text (left) and image (right) query tokens. MoL consistently produces lower entropy for image queries, indicating sharper attention. **Bottom:** cross-modal attention fractions. Image tokens attend more strongly to text keys under MoL (right), while text-to-image attention remains low for both models (left).